Prof. Philip Koopman

Positive Trust Balance

WAISE / September 2020
(Expanded Version)

Carnegie Mellon University

@PhilKoopman

EDGE CASE RESEARCH

Overview

EDGE CASE RESEARCH

- Positive Risk Balance (PRB):
  - Claim of risk lower than a human driver
  - But, how can you be sure this is true?

- Positive Trust Balance:
  - Claim that you can trust predicted PRB
  - Four components:
    - Validation ➔ Test it Right
    - Good engineering ➔ Build it Right
    - Field feedback ➔ Improve it Right
    - Strong safety culture ➔ Live it Right

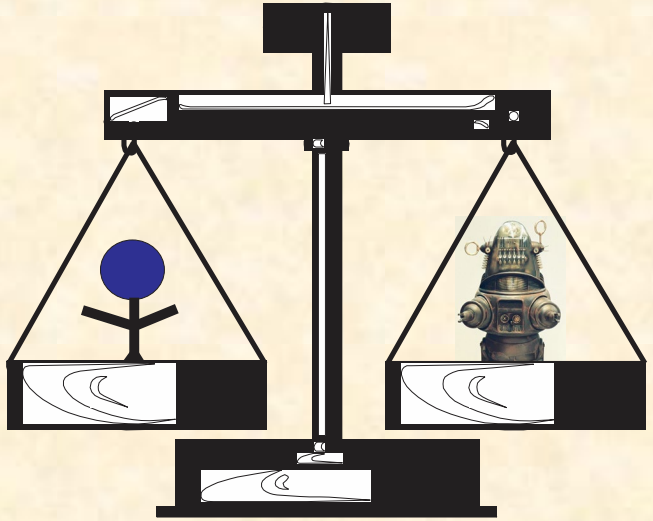
(Thanks to Wes Doonan for the four Rights)

Positive Risk Balance (PRB)

EDGE CASE RESEARCH

- PRB ➔ safer than human drivers
  - 1x?, 10x? or 100x? (other metric?)
- Brute force: drive a few billion miles
  - How often does safety driver intervene?
  - Invalidated if operational domain changes
  - Changing software resets odometer

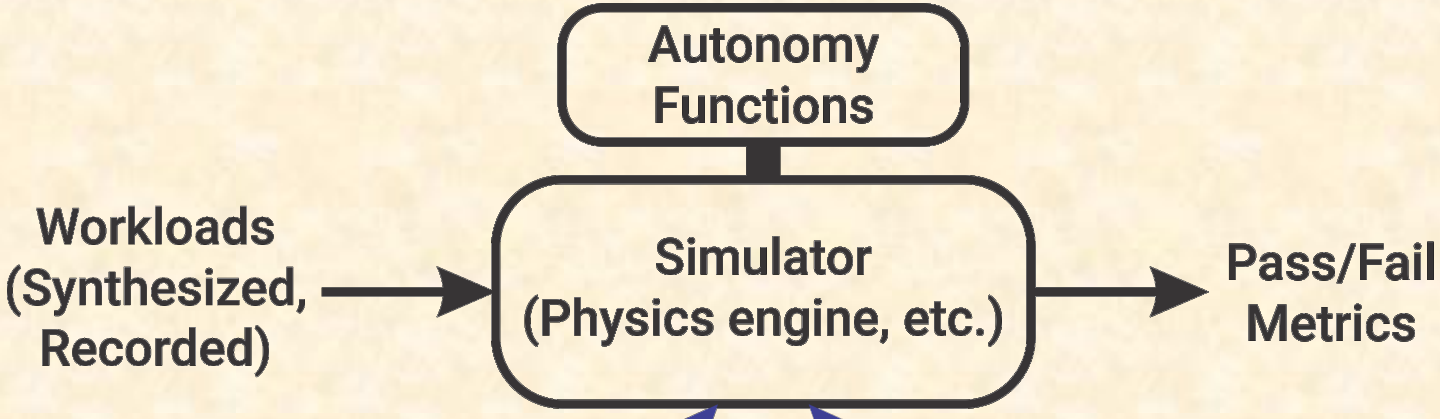
  … too expensive, takes too long
- Practical approach requires simulation



Simulation Validity

EDGE CASE RESEARCH

- All parts have to be valid to get a valid prediction



MODELS:
- Roads, weather, ...
- Sensors
- Vehicle dynamics
- Other road users
- Faults/failures
- What "safe" means ...

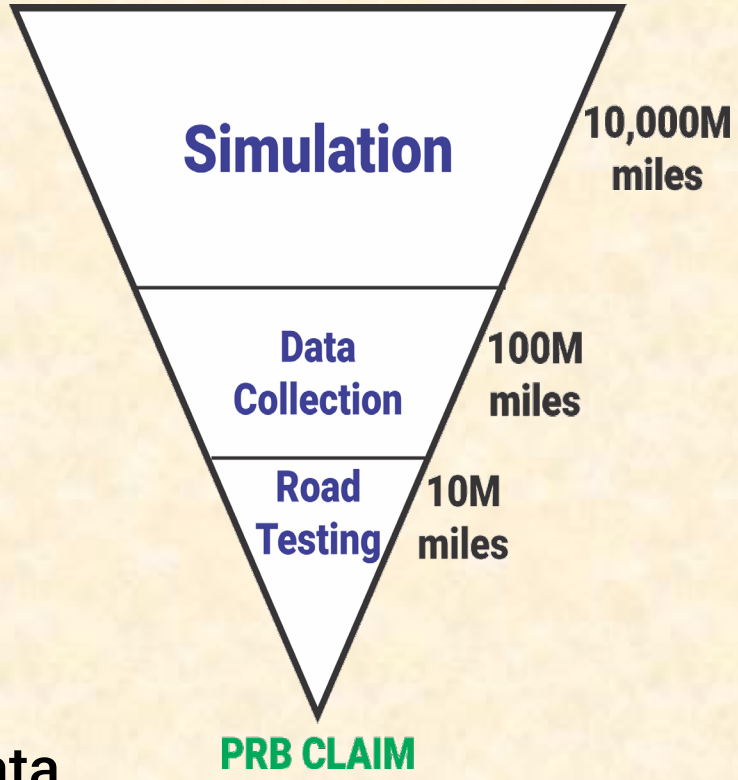
Design of Experiments

Does simulation approach include perception?

Hypothetical Validation Campaign

EDGE CASE RESEARCH

- 10,000M mile simulation campaign
  - Goal: under 1 fatality/billion miles
  - Claim ~10x PRB if simulation is valid
- 100M mile collected data/scenarios
  - Claim simulating this is representative
- 10M road testing of final software
  - Claim this validates simulation
- Is this statistically valid PRB?
  - Questionable confidence in collected data
  - Road testing useful, but insufficient on its own



How Much Do You Trust Simulation?

EDGE CASE RESEARCH

- Would put a child in front of a PRB self driving car?
  - 10,000M mile sims
    … perhaps with a simulator error?
  - 100M miles data collected
    … perhaps with scenario analysis errors?
  - 10M of road testing
    … that missed the above errors?
  - Built from software binaries
    … with no safety analysis?
  - With biased perception training data?

Claimed PRB Isn't Enough

EDGE CASE RESEARCH

- Validation-only "PRB" claim is really:

"We have PRB as far as we know"

BUT:

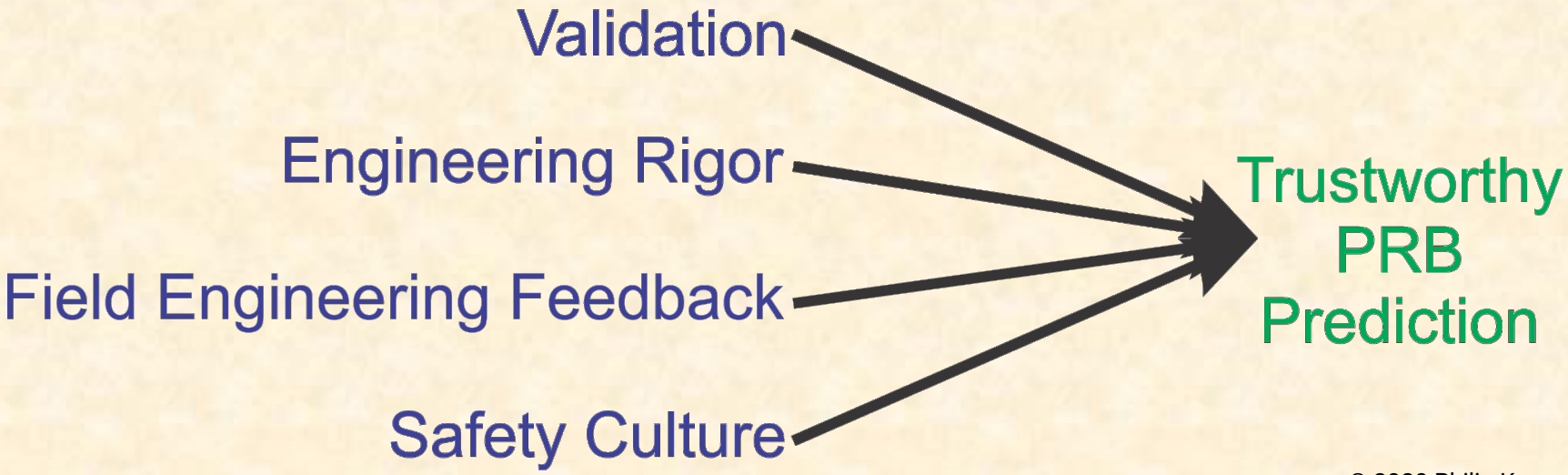
- Maybe we just got lucky that validation missed defects
- Maybe we missed something in our models
- Maybe we had confirmation bias due to time pressure

- Where's the safety argument?

Positive Trust Balance

EDGE CASE RESEARCH

- Stakeholders must trust that system is safe enough
  - Validation predicts PRB
  - Trust that PRB estimate is as valid as you can make it
  - Trust continuous improvement based on experience



Engineering Rigor

EDGE CASE RESEARCH

- Testing alone is insufficient for life-critical systems
  - So we use also use engineering rigor

- Can you trust the system itself?
  - Is it engineered for safety?
  - Were standards and best practices used?
  - Is there a safety case documenting all this?
- Can you trust your validation process?
  - Did you engineer the simulations properly?
  - Did you design the validation campaign properly?

Field Engineering Feedback

EDGE CASE RESEARCH

- Expected risk has a mean + uncertainty
  - You should deploy only when PRB mean is acceptable
  - But, there will be uncertainty
    - Missed edge cases during road testing
    - Unknown gaps in validation plan
    - Unknown unknowns in general

- Solution: continuous field monitoring
  - Monitor Safety Performance Indicators (SPIs)
    - SPI violation means safety argument has a defect
    - Investigate and fix root causes before loss events
  - Start during validation; continue after deployment

Safety Culture

EDGE CASE RESEARCH

- Did you do what you said you did?
  - Did your validation skip over known problems?
  - Did your engineering team skip process steps?
  - Is your field monitoring ignoring SPI violations?
- Good safety culture mitigates risk
  - Having a Safety Management System is ***a start***
  - Safety culture involves everyone in the lifecycle
- Safety culture simplified:
  - Are you incentivized to do the right thing?
  - Is it OK to tell your boss bad news? Will your boss fix it?

https://bit.ly/3i5wl57

Positive Trust Balance

EDGE CASE RESEARCH

- **Positive Trust Balance:**
  - Stakeholders trust that lifecycle risk will be acceptable (e.g., PRB)

EDGE CASE RESEARCH

WE DELIVER THE PROMISE OF AUTONOMY